PERGAMON

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Vision Research 43 (2003) 2265–2280

Vision Research

www.elsevier.com/locate/visres

Human visual object categorization can be described by models with low memory capacity

Robert J. Peters [a,*], Fabrizio Gabbiani [a,b], Christof Koch [a]

[a] *Computation and Neural Systems, Division of Biology, Caltech, 139-74 Pasadena, CA 91125, USA*
[b] *Division of Neuroscience, Baylor College of Medicine, Houston, TX 77030, USA*

Received 31 October 2001; received in revised form 20 August 2002

Abstract

Studies of high-level models of visual object categorization have left unresolved issues of neurobiological relevance, including how features are extracted from the image and the role played by memory capacity in categorization performance. We compared the ability of a comprehensive set of models to match the categorization performance of human observers while explicitly accounting for the models' numbers of free parameters. The most successful models did not require a large memory capacity, suggesting that a sparse, abstracted representation of category properties may underlie categorization performance. This type of representation—different from classical prototype abstraction—could also be extracted directly from two-dimensional images via a biologically plausible early-vision model, rather than relying on experimenter-imposed features.

Keywords: Visual object categorization; Exemplar model; Multidimensional scaling; Similarity; Representation; Pairs; Triads

1. Introduction

In humans, object categorization is one of the primary tasks of the visual system. Sensory processing of visual stimuli, along with prior visual experience, leads to categorization judgments that can ultimately be used for cognition. In the last 30 years, research in mathematical psychology has discovered much about the processes of visual categorization (e.g., Ashby, 1992a; Ashby & Maddox, 1993; Ashby & Waldron, 1999; Nosofsky, 1984, 1991; Reed, 1972; Smith & Minda, 1998) by combining the techniques of visual psychophysics and computational modeling to develop high-level theories of categorization. Despite the predictive success of these theories, there exists a gap between the descriptive framework of the models, and our current knowledge of the neuronal mechanisms involved in categorization. An important aim therefore is to shorten this gap by extending models so that their implementations are reasonable in light of recent developments in the neurophysiology of object recognition and categorization (Ashby & Ell, 2001; Freedman, Riesenhuber, Poggio, & Miller, 2001; Ishai, Ungerleider, Martin, Schouten, & Haxby, 1999; Kanwisher, McDermott, & Chun, 1997; Op de Beeck, Wagemans, & Vogels, 2001; Sigala & Logothetis, 2002). In this study, we address three key aspects of categorization models, each of which can be studied with psychophysical experiments and be informed by neurobiology.

First, current categorization models typically depend on high-level multidimensional representations of incoming stimuli (Ashby, 1992b, Chap. 16; Ashby & Maddox, 1993). Edelman (1999) reviewed evidence suggesting that such representations are intimately linked with the perceptual similarity of stimuli. A common technique used to infer implicit psychological representations is to apply multidimensional scaling (MDS) to observers' similarity judgments about a set of stimuli. Presently, the link between these psychophysical measures of similarity and the neuronal mechanisms underlying stimulus representation in the primate visual system remains poorly understood. New approaches using functional brain imaging in humans (Edelman, Grill-Spector, Kushnir, & Malach, 1998) and electrophysiological recordings in trained macaque monkeys (Op de Beeck et al., 2001; Sigala & Logothetis, 2002) are

* Corresponding author.
E-mail address: rjpeters@klab.caltech.edu (R.J. Peters).

0042-6989/$ - see front matter
doi:10.1016/S0042-6989(03)00279-7

likely to shed light on these issues. Such work will ultimately have to rely on comparisons between inferred psychological representations in monkey and human observers. Since it is nearly impossible to train animals to give graded similarity ratings between pairs of objects (the common method in human studies), animal studies must rely on two-alternative forced-choice methods instead. It is therefore important to directly compare these two ways of rating object similarity directly in human subjects.

Second, in a biological system, any high-level representation must be built from lower-level representations, and in vision this means that all representations must ultimately trace back to the retinal input. Many categorization models presuppose that the high-level (external) features used by the experimenter to define the objects are the same as those used internally by the observer when making a categorization decision. For example, many categorization studies have used a set of circles with bisecting lines, defined by two features: the diameter of the circle, and the angle of the bisecting line (e.g., Maddox & Ashby, 1993). This approach has certainly been fruitful, and MDS studies have demonstrated strong similarities between the external and internal feature representations. Nevertheless, apparent irregularities in the categorization process that might be inexplicable in terms of high-level representations, could appear entirely natural in the light of biological early vision. At the least, features such as *angle of the bisecting line* are not likely to be represented explicitly by neurons involved in visual perception; rather, a population of neurons might form a distributed representation, in which each neuron responds preferentially to a single range of orientations. Whether such differences have an effect on the output of categorization models is an empirical question. We have tested a set of hybrid models, in which an early-vision model based on Riesenhuber and Poggio (1999) is used to process the input in image space, yielding a set of coarse spatial maps, one for each of a small number of local image features. These maps are then used as input to the high-level categorization models after a dimensionality reduction step.

Third and last, the models approximate categorization decisions using a mechansim based on the multidimensional representation of incoming stimuli, plus possible auxiliary representations, such as memory traces. This process is typically controlled by a number of free parameters, which are fitted with the goal of matching human categorization behavior. However, a simple statistical comparison between models—even after accounting for the number of free parameters—may ignore important differences in the neurobiological implications of the models. For example, one successful model, the *generalized context model* (GCM; Nosofsky, 1984), assumes that all training images are stored in memory; a literal interpretation of the GCM might conclude that the neuronal substrate of categorization also scales linearly with the number of exemplars in a category, or that categorization in biological systems involves only simple memorization, without any category-level abstraction (Knowlton, 1999). To provide a more detailed look at such issues, we introduce a *roaming-exemplar model* (RXM) that draws from neural networks (Poggio & Girosi, 1990; Rosseel, 1996) and exemplar-based models of categorization (Kruschke, 1992; Nosofsky, 1991; Nosofsky, Kruschke, & McKinley, 1992). The RXM also has much in common with the *striatal pattern classifier* (SPC) of Ashby and Waldron (1999), including the fact that its memory traces are free parameters. This stands in contrast to previous exemplar-based models, and hence neurobiological plausibility can be assessed directly by accounting for numbers of free parameters when comparing fitted models.

2. Methods

2.1. Subjects

Eighteen psychophysics subjects (ages 18–25) from the Caltech community participated as paid volunteers in the experiments described below. Informed consent was obtained from all subjects, and experimental procedures were approved by the California Institute of Technology's Committee for the Protection of Human Subjects.

2.2. Stimuli

We used three types of schematic, line-drawn visual stimuli (Fig. 1): Brunswik faces and tropical fish outlines, which have been used previously (see below), plus a new set of "cartoon face" images. Each type of visual object was parameterized along four dimensions comprising the *stimulus parameter space*. Different sets of objects were assigned to *configurations*, which contained equal numbers of *training exemplars* assigned to each of two categories, as well as an additional number of *test exemplars*. The training exemplars from the two categories were always chosen so as to be linearly separable in the objects' parameter space; that is, the members of the two categories could be separated by some 3-D hyperplane in the 4-D parameter space.

2.2.1. Brunswik faces

These line-drawn face stimuli (Fig. 1a; Brunswik & Reiter, 1937) have been used frequently in categorization experiments both with human (Nosofsky, 1991; Reed, 1972) and non-human observers (pigeons, Huber & Lenz, 1996; monkeys, Sigala, Gabbiani, & Logothetis, 2002). Each face consists of a simple ovaloid outline with internal features defined by (compressed) circles

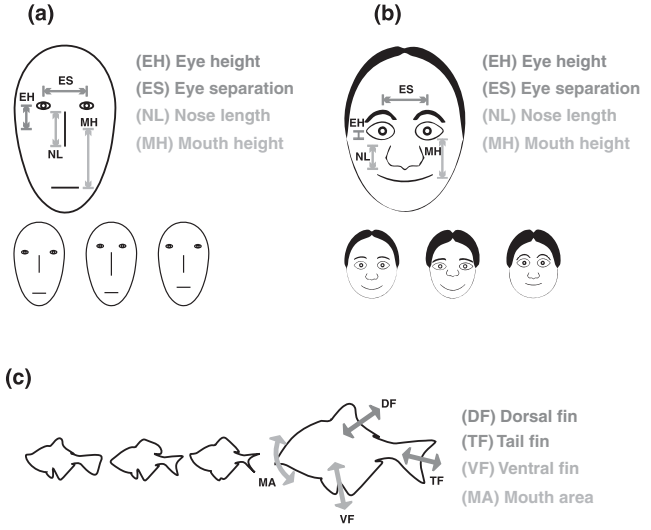

Fig. 1. Three object classes, each with four stimulus parameters controlling that object type, were used in similarity and categorization psychophysics tasks. Three sample objects of each type demonstrate the typical ranges of the parameters. (a) Brunswik faces. (b) Cartoon faces. Although these faces are described by 28 parameters, the present study used only the 4 parameters corresponding to those in (a). (c) Fish outlines.

and straight lines. The faces are parameterized by *eye height* (EH; the vertical distance from the centers of the eyes to the center of the face), *eye separation* (ES; the horizontal distance separating the centers of the eyes), *nose length* (NL; the vertical length of the nose line), and *mouth height* (MH; the vertical distance from the center of the face to the mouth line).

2.2.2. Cartoon faces

These stimuli (Fig. 1b) were introduced in an fMRI study (Jovicich, Peters, Koch, Chang, & Ernst, 2000) as a parameterized object type that produced stronger activation in the human fusiform face area (Kanwisher et al., 1997) than did Brunswik faces. The cartoon faces extend the Brunswik faces in several ways to make the faces appear more human: a simple band of hair is added around the top of the head, the size and dilation of the pupils may be varied, eyebrows are added above the eyes, the nose outline is defined by an extended open contour, and the mouth is defined as a Bezier curve rather than a straight line. To control these additional features, the cartoon faces have a total of 28 stimulus parameters; however, in the present study only the four parameters corresponding to the Brunswik face dimensions were varied, while the other 24 parameters were held constant.

2.2.3. Tropical fish outlines

These line-drawn images (Fig. 1c) were first used to offer a completely novel stimulus set to monkey observers in a categorization task (Sigala et al., 2002). Other fish images have been used previously in studies of categorization in people and pigeons (Hernstein & de Villiers, 1980) and in monkeys (Vogels, 1999). Each fish image is composed of four cubic spline curves that were fitted to scanned outlines of tropical fish. By adjusting one control point of each of the curves, four features of the outlines could be smoothly deformed: the dorsal fin (DF), tail fin (TF), ventral fin (VF), and mouth area (MA).

2.3. Similarity tasks

Two different similarity tasks (pairs and triads tasks) were performed with the 20-object configurations used in Experiment 1 (see Section 2.5; Fig. 2). For each configuration, subjects' psychophysical responses were used to form a 20×20 experimental *dissimilarity matrix* with entries δ_{ij}, using a procedure specific to the task (see descriptions below). This matrix was then used to estimate subjects' psychological representations of the stimuli (see Section 2.4).

In the pairwise comparison task (Borg & Groenen, 1997, Chap. 6.2) or *pairs task*, subjects viewed sequences

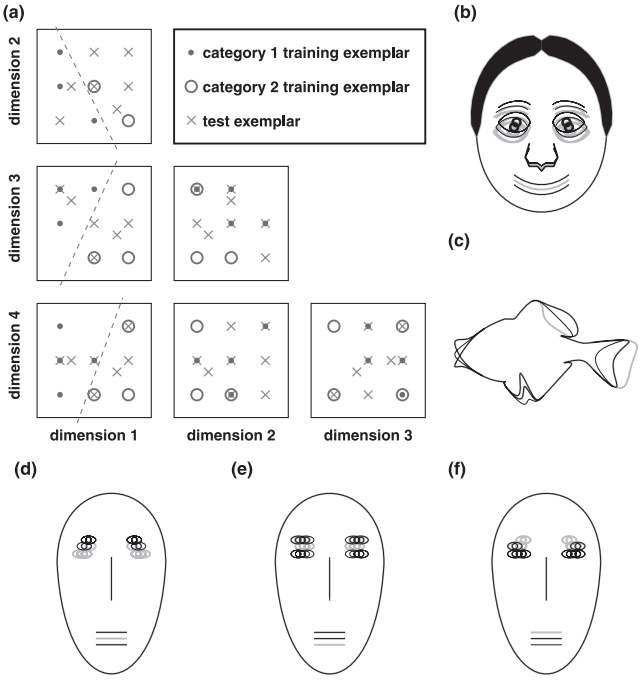

Fig. 2. Experiment 1 used five 20-object sets, each defined in a 4-D parameter space. (a) The abstract configuration is shown in projections onto the six possible pairs of dimensions. All-exemplars fall on a $3\times3\times3\times3$ grid, except for the two category prototypes, which were among the test exemplars. Dashed lines indicate where the two categories' training exemplars are linearly separable. (b–f) For illustration, the training exemplars of category one (thin black lines) are superimposed upon those of category two (thick gray lines), for (b) cartoon faces with dimensions {1 = EH, 2 = ES, 3 = NL, 4 = ML}, (c) fish outlines {TF, VF, DF, MA}, (d) Brunswik faces {EH, ES, NL, MH}, (e) Brunswik faces {NL, MH, EH, ES}, and (f) Brunswik faces {MH, EH, NL, ES}. See Fig. 1 for abbreviations.

of simultaneously presented pairs of objects. Each pair was presented for 2 s, followed by 2 s of blank screen. Subjects could respond at any time during that 4 s interval with a button press between 1 and 9, indicating how similar the objects appeared. Subjects were instructed to choose "9" if and only if the two objects were identical. Each of the 400 possible pairings of the 20 objects was presented three times throughout the experiment, giving 1200 total trials. For each pair of objects $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, the dissimilarity matrix entry δ_{ij} was taken to be $9 - \bar{s}_{ij}$, where $\bar{s}_{ij}$ is the average similarity rating over the n trials containing objects i and j ($n = 3$ for $i = j$, $n = 6$ for $i \neq j$).

The *triads task*, a variant of the anchor stimulus method (Borg & Groenen, 1997, Chap. 6.2), is a two-alternative forced-choice (2-AFC) task, and as such it has been particularly useful for studies involving non-verbal observers (e.g., human infants, Arabie, Kosslyn, & Nelson, 1975; monkeys, Sigala et al., 2002). Subjects viewed sequences of simultaneously presented triads of objects, arrayed horizontally. Each triad $(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3)$ was presented for 2 s, followed by 2 s of blank screen. Subjects could respond at any time during that 4 s trial with a button press indicating whether the left pair $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ or the right pair $(\boldsymbol{x}_2, \boldsymbol{x}_3)$ appeared more similar. Time constraints prohibited using all possible triad combinations. Instead, the 6840 possible triads $(\boldsymbol{x}_i, \boldsymbol{x}_j, \boldsymbol{x}_k)$ of the 20 objects were sorted by the Euclidean distance in stimulus parameter space between the leftmost and rightmost stimuli ($d(\boldsymbol{x}_i, \boldsymbol{x}_k)$), and the 1710 triads with the largest such distances were used for psychophysics. Finally, subjects' binary responses in the triads task were transformed into analog dissimilarities δ_{ij} using a procedure described in Sigala et al. (2002).

2.4. MDS analysis

Multidimensional scaling (MDS) was used to find a set $\widehat{\boldsymbol{X}} = \{\hat{\boldsymbol{x}}_1, \ldots, \hat{\boldsymbol{x}}_N\}$ of N 4-D vectors $\hat{\boldsymbol{x}}_i$, that best reflected the internal psychological representation used by a subject when performing a similarity task. The best such representation is found by minimizing the *stress*

$$\sigma = \frac{1}{2} \sum_{i,j} (d(\hat{\boldsymbol{x}}_i, \hat{\boldsymbol{x}}_j) - \delta_{ij})^2,$$

where d is the Euclidian distance and δ_{ij} are the dissimilarities computed from subjects' responses in one of the similarity tasks. [1] These representations allow for a clear correspondence between the scaled dimensions and the physical stimulus parameters, as explained next.

To align the MDS configuration $\widehat{\boldsymbol{X}}$ with the original configuration $\boldsymbol{X}$, we used an isometric *Procrustes transformation* P, consisting of a rigid rotation, translation, and uniform scaling (Borg & Groenen, 1997, Chap. 19). The optimal Procrustes transformation $P_{\min}$ minimizes the loss function $L(P) = \sum_i d^2(\boldsymbol{x}_i, P(\hat{\boldsymbol{x}}_i))$. This minimum value $L(P_{\min})$—the *residual squared distance* (RSD)—quantifies the dissimilarity between subjects' psychological representation $\widehat{\boldsymbol{X}}$ and the original stimulus configuration $\boldsymbol{X}$.

To determine whether the observed RSDs were smaller than would be expected by chance, a Monte Carlo technique was used. RSDs were computed between the original configuration and 10^5 random configurations whose parameters were drawn from a uniform distribution over $[0, 1]$. The resulting distribution was used to estimate the significance levels of the RSDs of the pairs and triads MDS configurations.

2.5. Categorization tasks

The categorization experiments consisted of a training phase and a testing phase. In both phases, subjects viewed a series of objects presented one at a time. Each object was presented for 2 s, followed by 2 s of blank screen. During each 4 s trial, subjects pressed one of two buttons indicating to which category the object belonged. In the training phase, subjects were shown only the two categories' training exemplars, and were given feedback in the form of a high- or low-pitch tone indicating whether their response was correct or incorrect, respectively. Subjects performed training blocks of 100 trials until they scored $\geqslant 85\%$ correct on a single block. Next, they moved into the testing phase, in which they were shown the previously unseen test exemplars in addition to the training exemplars that they had viewed during the training phase. Subjects received no feedback on their responses during the testing phase.

In Experiment 1, the values for each stimulus dimension were quantized to three possible values for each dimension, so that the set of possible objects lay on a $3 \times 3 \times 3 \times 3$ grid in stimulus parameter space. The configuration of 20 objects on this grid (Fig. 2a) followed that used in Nosofsky (1991) and Sigala et al. (2002), with 5 training exemplars for each category, plus 10 test exemplars that included the two category prototypes. For each set of objects, each of the four stimulus parameters for that object type was assigned to one of the four generic dimensions in the stimulus configuration shown in Fig. 2a. It is significant how the parameters are assigned, since each generic dimension carries different information about category membership. For example, the categories were linearly separable in projections onto 2-D planes for pairs of stimulus dimensions $(1, 2)$, $(1, 3)$,

[1] Note that this procedure deviates from a strict definition of MDS because the dimensionality of the representation space was fixed to 4, rather than being a free parameter. However, previous studies using Brunswik faces and fish stimuli have obtained satisfactory MDS solutions with 4-D representations (Nosofsky, 1991; Sigala et al., 2002).

and (1,4), so dimension 1 was more informative about an object's category than were the other dimensions. In all, five sets of stimuli were used in Experiment 1. These included three sets of Brunswik faces in which the stimulus parameters were assigned to the generic dimensions in different orderings ({EH, ES, NL, MH}, {NL, MH, EH, ES}, and {MH, EH, NL, ES}), a set of cartoon faces ({EH, ES, NL, ML}), and a set of fish outlines ({TF, VF, DF, MA}).

In Experiment 2, a larger configuration of 80 objects was used (Fig. 3), with 10 training exemplars for each of the two categories, plus 60 test exemplars. The exemplars were arranged on a $7\times7\times7\times7$ grid in the stimulus parameter space. There were 12 such sets, identical except that the discretization grid of each set was rotated through different angles ($\theta = n \cdot 15°$, $n \in [0,\ldots,11]$) in the eye-height/eye-separation plane of parameter space.

2.6. Categorization models

We tested several categorization models by fitting them to match the human observers' response profiles from the testing phase of the categorization tasks. Each model receives input in a 4-D feature space (i.e., not image space), and produces an output that represents a categorization probability for the input object. The models we tested fall into several categories, each of which proposes a unique architecture for the categorization process (see Fig. 4), with different free parameters, and different assumptions about the memory usage of the system being modeled. These factors must be weighed along with the raw goodness-of-fit when assessing the neurobiological plausibility of the different models.

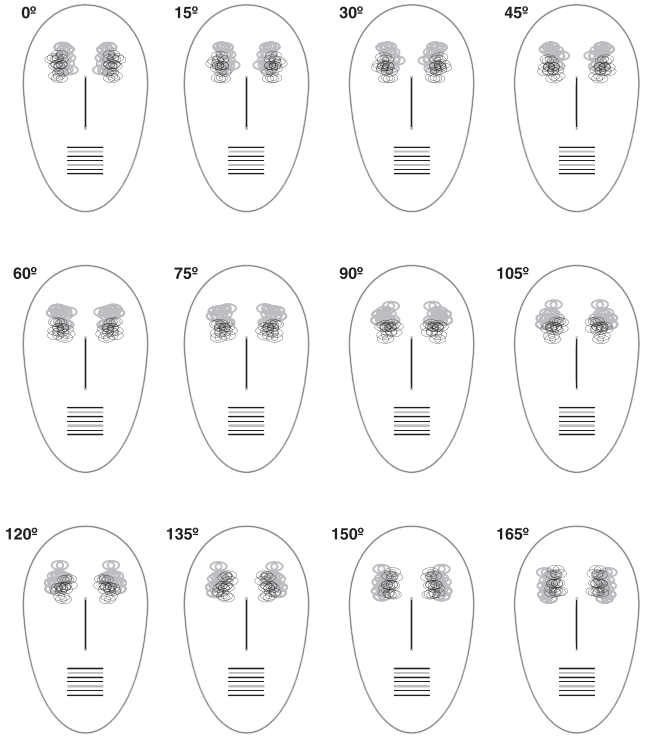

Fig. 3. Experiment 2 used these 12 sets of Brunswik faces. Each image shows the 10 training exemplars of category one (thin black lines) superimposed upon the 10 training exemplars of category two (thick gray lines). The sets differ only in the angle by which the objects are rotated in eye-height/eye-separation plane of feature space.

In general, we assume (1) that each exemplar is described by a point in an R-dimensional space (Ashby, 1992b, Chap. 16), $\boldsymbol{x} = (x_1,\ldots,x_R)$, whose components may be drawn either from the original stimulus configuration, from an MDS configuration, or from a configuration based on features extracted from an early-vision model (see Section 2.6.5), and (2) that each category is defined by N training exemplars $\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N\}$.

2.6.1. Exemplar models

Exemplar models associate memory traces of M ($1 \leqslant M \leqslant N$) stored exemplars [2] $\{\boldsymbol{y}_1,\ldots,\boldsymbol{y}_M\}$ with each category. Several model subtypes differ in the way that these stored exemplars are selected:

- *All-exemplar models* (Fig. 4a) assume $M = N$, and $\boldsymbol{y}_i = \boldsymbol{x}_i$. All of the training exemplars are explicitly stored in memory, so these models have a high memory demand that is linear in the number of training exemplars. All-exemplar models include the *average-distance model* (ADM, Reed, 1972) and *generalized context model* (GCM; Nosofsky, 1991).
- *Prototype (one-exemplar) models* (Fig. 4b) assume $M = 1$; each category stores only the arithmetic mean of the category's training exemplars, $\boldsymbol{y}_1 = 1/N \sum_i \boldsymbol{x}_i$. These models have low and constant memory demand, independent of the number of training exemplars; however, the models imply a more complex computational mechanism to estimate the prototype during trial-by-trial exposure to the training exemplars. Prototype models include the *weighted prototype model* (WPM; Reed, 1972) and the *weighted prototype similarity model* (WPSM; Nosofsky, 1991).
- In the proposed *roaming-exemplar model*$\langle M\rangle$ (RXM$\langle M\rangle$, Fig. 4c), each category stores M exemplars, each of which is a linear combination of the training exemplars for that category, $\boldsymbol{y}_j = \sum_i w_{ij}\boldsymbol{x}_i$. Under the neurobiological consideration that neurons do not represent objects far different from those

[2] Our usage of the term "exemplar" to denote stored memory traces reflects a meaning of *ideal meaning or pattern* or *prototype*, rather than a strict meaning of *previously seen stimulus*. For example, in the RXM, the stored exemplars are generalizations of the memory traces used in all-exemplar or prototype models, and are most likely not previously seen stimuli.

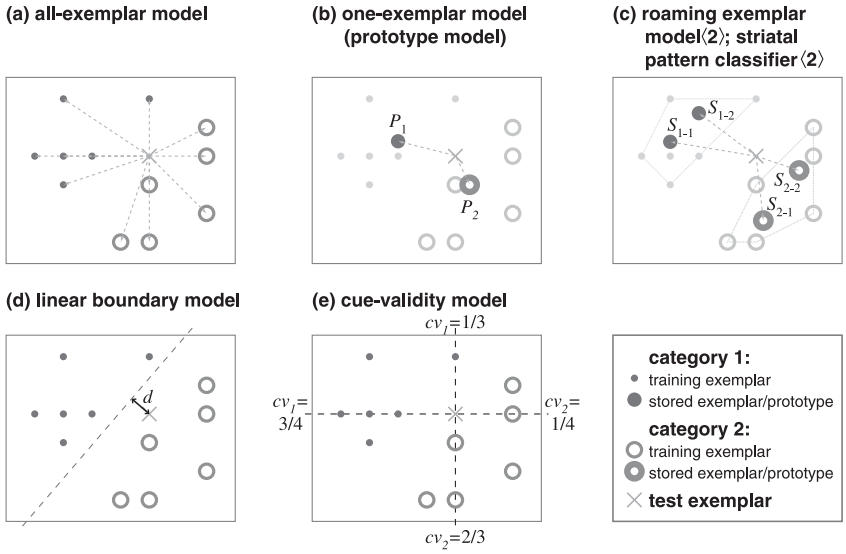

Fig. 4. Schematic depictions of several kinds of categorization models. Each diagram shows a hypothetical set of training exemplars from two categories (● and ○) in a 2-D feature space, plus a test exemplar (×) which is to be classified. (a–c) Three types of models which rely on distances (indicated by dashed lines) between a test exemplar and each stored exemplar from both categories: (a) *all-exemplar model*, in which the set of stored exemplars is just the set of training exemplars; (b) *one-exemplar*, or *prototype model*, in which the single stored exemplar per category is the arithmetic mean of that category's training exemplars; (c) *roaming-exemplar model*$\langle M\rangle$ (RXM$\langle M\rangle$) and *striatal pattern classifier*$\langle M\rangle$ (SPC$\langle M\rangle$), in which each category has M (in this case, $M = 2$) stored exemplars, which must lie within the polygon that circumscribes the training exemplars (dotted lines). The RXM$\langle M\rangle$ uses a summed-similarity decision rule, while the SPC$\langle M\rangle$ uses a uses a nearest-neighbor decision rule. (d) *Linear boundary model*, in which the model uses a linear boundary that separates the categories to classify test exemplars according to the side of this boundary on which they fall. (e) *Cue-validity model*, which classifies a test exemplar according to the total cue-validity across all features; the cue-validity cv_i for category i of a given feature is the posterior probability of an exemplar with that feature belonging to category i (values of cv_1 and cv_2 are shown).

that have been previously observed, the stored exemplars are restricted to a region circumscribed by the training exemplars, so the weights are constrained by $w_{ij} \geqslant 0$ and $\sum_i w_{ij} = 1$ for all j. The number of stored exemplars M is *not* a free parameter of a given RXM$\langle M\rangle$, but the stimulus parameters of those stored exemplars *are* free parameters of the model. Thus, when the RXM is fitted to a dataset, the number of stored exemplars is chosen and fixed at the start, although RXM$\langle M\rangle$'s with different (fixed) values of M may be fitted to the same dataset. The memory demand of the RXM$\langle M\rangle$ varies between that of the prototype models (for $M = 1$) and that of the all-exemplar models (for $M = N$); the computational complexity is similar to that of the prototype models, since some mechanism must adjust the stored exemplars during training.

Next, the exemplar model computes a similarity measure between the test exemplar $\boldsymbol{x}$ and each of the stored exemplars $\boldsymbol{y}$, based on a weighted Euclidean distance: $d_{\boldsymbol{\alpha}}(\boldsymbol{x},\boldsymbol{y}) = \left(\sum_j \alpha_j (x_j - y_j)^2\right)^{1/2}$, with $\alpha_j \geqslant 0$ and $\sum \alpha_j = 1$ (other metrics are possible; see e.g., Ashby & Maddox, 1993). The coefficients α_j, called *attentional weights*, are intended to model the ability of human observers to attend preferentially to the most task-relevant stimulus features. The similarity s decays with the distance d, either linearly ($s = -d$, as in the RXM, ADM, and WPM), or exponentially ($s = \mathrm{e}^{-cd}$, as in the GCM and WPSM; see Shepard, 1987).

Then, for each test exemplar $\boldsymbol{x}$, the evidence E_i for category C_i is given as the sum of similarities between $\boldsymbol{x}$ and the M stored exemplars $\boldsymbol{y}_j^i$ of that category: $E_i(\boldsymbol{x}) = \sum_{j=1}^{M} s(\boldsymbol{x}, \boldsymbol{y}_j^i)$. Finally, the model's categorization of $\boldsymbol{x}$ is based on the expression $E_1(\boldsymbol{x}) - E_2(\boldsymbol{x}) + n > t$, where n represents zero-mean Gaussian noise with variance σ^2, and t is a threshold parameter; $\boldsymbol{x}$ is assigned to category 1 if this expression is true, otherwise to category 2.

The free parameters of the exemplar models are thus $(\boldsymbol{\alpha}, c, t, \sigma)$, plus $2M$ stored exemplars for the RXM$\langle M\rangle$.

2.6.2. Striatal pattern classifier

The RXM shares a very similar mathematical formulation with the *striatal pattern classifier* (SPC) proposed by Ashby and Waldron (1999), although the mathematical elements have been treated with different neurobiological interpretations (Ashby & Ell, 2001). Both kinds of model rely on a set of units that represent different locations in feature space, but the models differ in how each category's evidence is computed for a given test exemplar. The exemplar models compute the sum of similarities between the test exemplar and each stored exemplar, whereas the SPC associates a test exemplar with the category of the nearest striatal pattern (in this respect the SPC resembles a k-nearest neighbor model

with $k = 1$). Both the SPC and the RXM use a similarity measure that decays linearly with distance. In order to maintain a formal similarity with the other models, we used the following decision rule for the SPC: for each test exemplar, the evidence for each category is given by the maximum of the similarities between the test exemplar and that category's stored exemplars. Thus, in the case of one stored exemplar per category, the SPC$\langle 1\rangle$ and the RXM$\langle 1\rangle$ form identical decision surfaces. However, with $M > 1$, the SPC$\langle M\rangle$ has a piecewise-linear boundary, while the RXM$\langle M\rangle$ has a curved decision boundary.

2.6.3. Boundary models

Decision bound theory (Ashby & Maddox, 1993) proposes that human perceptions of category exemplars are instances of random variables with multivariate normal distributions. Given a particular perception, the optimal decision strategy is to choose the category of which that perception was more likely an instance. Thus the decision boundary (the locus where both categories have equal probability densities) falls along the intersection of the graphs of the two probability density surfaces. If the covariance matrices of the exemplar distributions are identical for the two categories, then the decision boundary is a linear surface (i.e., a hyperplane); otherwise, it is a quadratic surface.

We tested the probit linear model (PBI; Fig. 4d; Ashby & Gott, 1988), which is trained to separate the categories' training exemplars with a boundary described by a normal vector $\boldsymbol{b}$ and a threshold t. Following training, a test exemplar $\boldsymbol{x}$ is classified according to the side of the boundary on which it falls:

$$\boldsymbol{x}\cdot\boldsymbol{b} + n > t \Rightarrow \boldsymbol{x} \in C_1.$$

The PBI model parameters are $(\boldsymbol{b}, t, \sigma)$; however the variance of the noise is assumed to be $\sigma^2 = 1$, since identical models are obtained for $(\boldsymbol{b}, t, \sigma)$ as for $(\lambda\boldsymbol{b}, \lambda t, \lambda\sigma)$ with $\lambda \neq 0$.

2.6.4. Cue-validity models

Cue-validity models (Fig. 4e) treat each stimulus parameter as an independent indicator of category membership, based on the relative numbers of exemplars from the two opposing categories that exhibit the *cue* (a particular value of a stimulus parameter). Thus, for example, a beard is a somewhat uncommon feature of male faces, yet it is an even less common feature of female faces, and so provides a highly valid cue to the gender category of a face.

In the weighted cue-validity model (WCVM; Reed, 1972), the validity for category C_i of the jth parameter x_j of a test exemplar $\boldsymbol{x}$ is defined as $v_{ij}(\boldsymbol{x}) = p(C_i|x_j)$. The overall cue-validity V_i is a weighted sum of these validities, $V_i(\boldsymbol{x}) = \sum_j \alpha_j v_{ij}(\boldsymbol{x})$, where the α_j are attentional weights as in the exemplar models, with $\alpha_j \geqslant 0$ and $\sum_j \alpha_i = 1$. Also as in the exemplar models, the decision rule incorporates Gaussian noise n and a threshold t; if the expression $V_1(\boldsymbol{x}) - V_2(\boldsymbol{x}) + n > t$ is true, $\boldsymbol{x}$ is assigned to category 1, otherwise to category 2.

A modified version of this model, called the weighted frequency cue-validity model (WFCVM; Reed, 1972), uses a different definition for the validity. A weight factor, $q = (1 + F(x_m))^{-1}$, is computed from the overall number of times $F(x_m)$ that the parameter value x_m occurs in exemplars from both categories. Then the WCVM's original validity v_{ij} is used to define the new validity $\tilde{v}_{ij}(\boldsymbol{x}) = \frac{1}{2}\cdot q + v_{ij}(\boldsymbol{x})\cdot(1-q)$, so that the validities of rare parameter values carry little information about category membership. This reflects the idea that subjects will pay more attention to common features.

The free parameters for both the WCVM and the WFCVM are $(\boldsymbol{\alpha}, t, \sigma)$.

2.6.5. HMAX

In order to assess the biological plausibility of the categorization models from a computational perspective, we adapted a hierarchical model of early vision ("HMAX") presented by Riesenhuber and Poggio (1999). HMAX operates directly in image space, in contrast to the categorization models described above, which operate in feature space. Our approach was to extract a new feature space representation from the output of HMAX, which could then be used as an alternate input for fitting the categorization models, to be compared with model fits obtained using the original physical feature space.

In brief, HMAX operates through two stages of "simple" and "complex" units (S1, C1, S2, and C2). The S1 representation is obtained by filtering the image with a bank of Gabor-like filters tuned for multiple orientations and spatial scales. The C1 representation is produced by pooling the activations of S1 units at neighboring spatial locations and across similar spatial scales. At the S2 level, more complex features are formed by pooling the activations of a 2×2 spatial array of neighboring C1 units tuned to specific orientations; in this way, different S2 units begin to represent features such as "elongated contour" or "corner" or "disk". Finally, each C2 unit pools across S2 units tuned to the same feature type, but at different spatial scales and/or spatial locations.

We made several modifications relative to the original model of Riesenhuber and Poggio (1999); these modifications were guided by the goal of increasing the variance of the HMAX outputs across the set of input images, so as to provide a rich but compact foundation for a subsequent categorization stage. First, instead of each C2 unit pooling across the entire image space, we subdivided the image into a 6×6 grid, with each C2 unit responding only to one of the 36 subregions. This increased granularity allowed the model to extract features

that were more relevant as input to the categorization models. In addition, we restricted the number of orientation filters among the S1 units from four to two (i.e., just horizontal and vertical). This retained the model's ability to represent the variability among the simple schematic input images, but at the same time significantly reduced the dimensionality of the output space: since each S2/C2 feature type represented a four-part configuration of two possible S1/C1 orientations, there were $2^4 = 16$ S2/C2 feature types (rather than $4^4 = 256$ as in the original model). With 36 spatial locations, this gave a total of $36 \times 16 = 576$ C2 units. To reduce this representation to a manageable size for input to the categorization models, we applied principal component analysis to the C2 activation vectors obtained across all of the input images in a set. In general, we found that >95% of the variance could be recovered with the first 50 of the 576 principal components, and ≈80% of the variance was recovered with only the first 4 components. Therefore, for comparison with the four-dimensional physical parameter configurations, we used the first 4 principal components from the modified-HMAX C2 activations to test how well the categorization models would fare with a biologically plausible input derived from the image space representation of the stimuli.

2.7. Model-fitting

We fitted models based on (1) the objects' physical parameter values, (2) the psychophysical (i.e., MDS) parameter values obtained from a pairs or triads task, and (3) the features derived via PCA from the C2 activations of the HMAX model. Furthermore, each model could either be fitted separately to each individual subjects' data, or be fitted once to data pooled across subjects. However, since pooled fits may not accurately reflect the categorization processes of individual observers (Maddox, 1999), we used only models fitted to individual subjects' data.

Each model's free parameters were fitted to maximize its ability to predict the categorization probabilities obtained from human observers. The goodness of this fit was quantified with the *likelihood*, L, of the model having generated the observed probabilities, given that the fitted model correctly describes the subject's categorization process (Collett, 1991). This likelihood is the conditional probability of the set of observed probabilities p_i, given the values of the model parameters (which govern the predicted probabilities $\widehat{p_i}$), over the N stimulus objects:

$$L = \prod_{i=1}^{N} \binom{n_i}{p_i n_i} (\widehat{p_i})^{p_i n_i} (1 - \widehat{p_i})^{(1-p_i)n_i},$$

where n_i is the number of categorization trials performed for object i, and $p_i n_i$ is the number of trials in which the observer assigned object i to category one. The likelihood takes the form of a binomial distribution because subjects' responses are treated as independent binary random variables. A numerical implementation of adaptive simulated annealing (Ingber, 1989) followed by a simplex method (Nelder & Mead, 1965) was used to maximize the likelihood L, or equivalently, minimize the minus *loglikelihood* ($-\ln L$), which can be computed more efficiently. The range of the likelihood is $0 \leqslant L \leqslant 1$, so the range of the minus loglikelihood is $\infty \geqslant -\ln L \geqslant 0$.

We used the percentage of variance (%-variance) explained by the model as a more tangible measure for comparing fitted models. This measure is simply given by r^2, the square of the correlation coefficient between the observed and predicted probabilities.

Finally, although the loglikelihood ($\ln L$) or %-variance are appropriate statistics for comparing fitted models having similar numbers of free parameters, comparisons of models differing in their number of free parameters, N_{fp}, require a statistic such as the Akaike information criterion (Zucchini, 2000), $\mathrm{AIC} = -2\ln L + 2N_{\mathrm{fp}}$, which contains a penalty term proportional to N_{fp}. Pairwise model comparisons were made with the Wilcoxon signed-rank test of either $-\ln L$ or the AIC, and we report the median value of $-\ln L$ or the AIC to summarize the model fits from a group of individual subjects.

3. Results

3.1. MDS fits

In order to quantify the goodness-of-fit between subjects' Procrustes-transformed MDS configurations and the original stimulus configuration, we used Monte Carlo simulations comparing the residual squared distances (RSDs) of our subjects' MDS configurations with the RSDs of random configurations (see Fig. 5a). The mean of the pairs-MDS distribution (0.1444) was roughly twice as close to the original configuration as would be expected by chance (0.3268), and all pairs-MDS configurations were significantly closer ($p < 0.005$) to the original configuration than were the random configurations. Likewise, the triads-MDS configurations were also all significantly closer to the original space than would be expected by chance ($p < 0.005$), although the mean of the triads-MDS distribution (0.1982) was not as close to the original configuration as was the pairs-MDS distribution. A paired t-test showed that the residual squared distances of the pairs-MDS configurations were significantly smaller than those of the triads-MDS configurations ($p < 0.05$).

To further assess the relationship between these two methods for obtaining similarity judgments, we per-

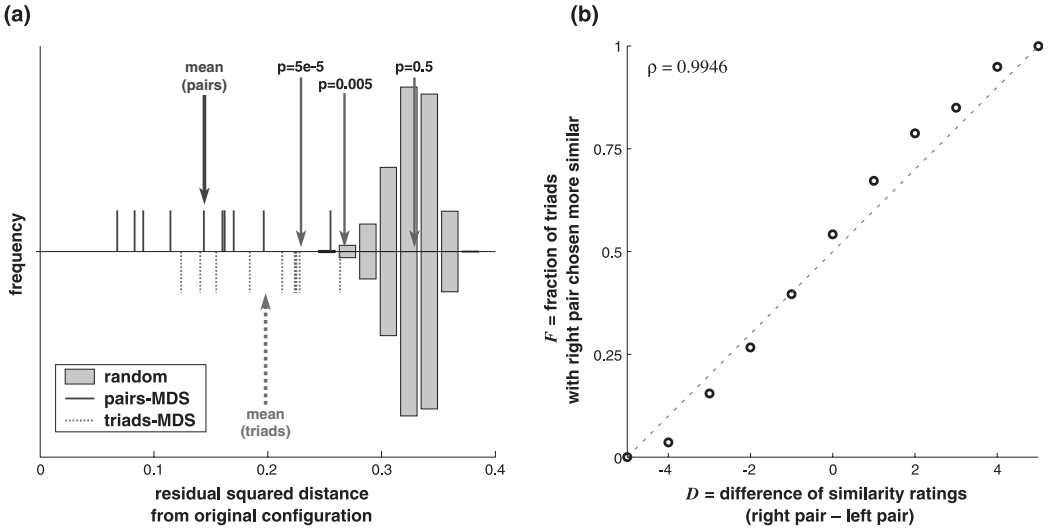

Fig. 5. A summary of the MDS configurations obtained with pairs and triads similarity tasks. (a) As measured with the *residual squared distance* (RSD), all of the pairs-MDS and triads-MDS configurations were significantly more similar ($p < 0.005$) to the original configuration of stimulus parameter values than would be expected by chance. The distribution of RSDs for 10^5 random configurations (gray bars, arrows with p-values) was compared with the RSDs for 10 subjects' pairs-MDS (upper, solid lines) and triads-MDS (lower, dashed lines) configurations. Two identical configurations would give an RSD of 0, while two unrelated configurations would give an RSD near the median of the random distribution (0.33). The RSDs for pairs-MDS were significantly smaller than those for triads-MDS ($p < 0.05$). (b) To directly compare the similarity judgments obtained in the pairs and triads tasks, we computed two metrics for triads of objects $(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3)$: (1) the difference $D = S(\boldsymbol{x}_2, \boldsymbol{x}_3) - S(\boldsymbol{x}_1, \boldsymbol{x}_2)$ of two similarity ratings given *in the pairs task*, and (2) among triads with similar values of D, the fraction F of trials in which the observer chose $(\boldsymbol{x}_2, \boldsymbol{x}_3)$ as more similar than $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ when viewing $(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3)$ *in the triads task*. The two measures D and F were highly correlated ($\rho = 0.9946$) across 10 subjects.

formed a more direct comparison, using subjects' raw responses rather than the derived MDS configurations (Fig. 5b). In each trial in the triads task, subjects viewed three objects $(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3)$ and compared the similarities of the two pairs $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ and $(\boldsymbol{x}_2, \boldsymbol{x}_3)$. Subjects also directly rated the similarities of these pairs in the pairs task. Thus, for each triad $(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3)$ which was shown in the triads task, we computed D_{pairs}, the difference between the similarity ratings given by the subjects *in the pairs task* to the pairs $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ and $(\boldsymbol{x}_2, \boldsymbol{x}_3)$. We then split the triads trials into groups with similar values of D_{pairs}. Within each group we computed F_{triads}, the fraction of trials for which the subject chose the right pair as more similar than the left pair *in the triads task*. These two measures D_{pairs} and F_{triads} were highly correlated in data obtained from single subjects ($\rho > 0.98$ for 9 of 10 subjects) and when data were pooled across subjects ($\rho = 0.9946$; Fig. 5b).

3.2. Model fits

We found no systematic differences in the fits obtained from different model subtypes (such as those using exponential vs. linear decay of similarity with distance). Therefore, in further discussion, models are referred to by their general names (e.g., all-exemplar models) rather than by the specific subtypes (e.g., ADM or GCM).

3.2.1. Model fits: Experiment 1

Table 1 summarizes the fits of the all-exemplar, linear boundary, prototype, and cue-validity models, for each of the five sets of objects used in Experiment 1, along with significance values for pairwise comparisons of the models using the Wilcoxon matched pair signed-rank test. [3] There were two general patterns of model fits.

The first pattern was associated with the first two Brunswik face sets ({EH, ES, NL, MH} and {NL, MH, EH, ES}, which depend primarily on attention to the eyes and nose) and the cartoon faces ({EH, ES, NL, ML}). In this pattern, the all-exemplar models obtained the best fit, but the boundary model also fit well, indistinguishable from the exemplar models. The prototype models fit significantly worse ($p < 0.05$) than the all-exemplar models, but the magnitude of this difference was small. Finally, the cue-validity models fit significantly worse than the other models.

The second pattern was seen with the third Brunswik face set ({MH, EH, NL, ES}) and the fish outlines ({TF, VF, DF, MA}). As in the first pattern, the all-exemplar models obtained the best fit. However, the rest of the

[3] Note that the RXM and SPC were not used in fitting the data from Experiment 1 because even with one stored exemplar, these models carry almost as many free parameters as the number of data points to be fitted (20). This renders any comparisons among such models virtually meaningless. This issue is avoided in Experiment 2 due to the greater number of test exemplars (80).

Table 1
Goodness-of-fit of the models tested in Experiment 1

		GCM	PBI	WPSM	WCVM
Brunswik faces {EH, ES, NL, MH}	%-variance	**98.22**	98.08	96.39	88.37
	$-\ln L$	**21.15**	22.23	27.32*	42.50*
Brunswik faces {NL, MH, EH, ES}	%-variance	95.68	**97.75**	95.32	74.38
	$-\ln L$	28.08	**26.53**	32.81	42.58*
Brunswik faces {MH, EH, NL, ES}	%-variance	**94.02**	58.56	61.55	86.30
	$-\ln L$	**36.83**	80.57*	90.31*	52.76
Cartoon faces	%-variance	**95.50**	90.70	90.18	86.66
	$-\ln L$	30.68	**29.95**	37.07*	53.49
Fish outlines	%-variance	**97.23**	80.98	70.30	96.03
	$-\ln L$	**20.73**	32.85*	74.36*	28.74

%-variance: larger value indicated better fit.
$-\ln L$, minus loglikelihood: smaller value indicates better fit.
Bold numbers: model(s) which gave the best fit in each row.
*Models whose $-\ln L$ was significantly worse ($p < 0.05$) than the best-fitting model in each row.

pattern was qualitatively different from the first pattern. Whereas the cue-validity models gave the worst fits in the first pattern, their fits were indistinguishable from the all-exemplar models in the second pattern. In addition, the boundary model fit very poorly, significantly worse than the exemplar models ($p < 0.05$). Finally, the prototype models fit even more poorly, significantly worse than the exemplar and boundary models ($p < 0.05$).

Each of the models tested in Experiment 1 was also fitted using MDS-based configurations obtained from the pairs or triads tasks. Measured by %-variance, both the MDS-pairs and MDS-triads model fits were strongly correlated with the fits obtained using the original configuration, as well as with each other ($\rho > 0.90$ in each case). The average goodness-of-fit of the MDS-pairs models lagged behind that of the original models by 2.3 %-variance, and the MDS-triads models lagged by an additional 5.5 %-variance.

3.2.2. Model fits: Experiment 2

We fitted subjects' categorization probabilities from Experiment 2 with versions of the roaming-exemplar model and striatal pattern classifier using 1, 2, 3, 5, 7, and 10 stored exemplars, [4] as well as the all-exemplar, prototype, and linear boundary models, and assessed these fits with three measures (see Table 2): the loglikelihood, the %-variance explained, and the Akaike information criterion (AIC).

When the model fits were assessed with their minus loglikelihoods (Table 2, row 2), we observed a pattern among the previously tested models similar to the first pattern observed in Experiment 1: the all-exemplar and boundary models both obtained better (lower) scores than the prototype model. However, each of these previous models was outperformed by all versions of the roaming-exemplar model and striatal pattern classifier. In addition, for both the RXM$\langle n \rangle$ and the SPC$\langle n \rangle$ the goodness-of-fit increased with the number n of stored exemplars—an unsurprising result, given that each stored exemplar reflects additional free parameters. The %-variance values (Table 2, row 1) show a similar pattern, but give a more concrete assessment of how well the models match the human subjects' categorization behavior: the best-fitting model (the SPC$\langle 3 \rangle$) captured nearly 92% of the variance, while the worst-fitting model (the WPSM) captured roughly 85% of the variance.

In contrast, when the model fits were assessed with the AIC to account for their numbers of free parameters (Table 2, row 3), the RXM and SPC with one stored exemplar per category (RXM$\langle 1 \rangle$ and SPC$\langle 1 \rangle$) obtained the best (lowest) scores among all models. These comparisons were statistically significant (Wilcoxon signed-rank test, $p < 0.05$) except against the PBI ($p = 0.44$). Moreover, increasing the number of stored exemplars in either the RXM$\langle n \rangle$ or SPC$\langle n \rangle$ was detrimental to the AIC goodness-of-fit; the SPC$\langle 10 \rangle$ (AIC = 253.29) and RXM$\langle 10 \rangle$ (SPC = 271.85) fit much worse than any of the other models.

Each of the models was also fitted using representations of the visual objects based on features derived from the C2 activations of the HMAX model, rather than the original physical parameters of the stimuli. We found that the features derived from HMAX recovered much of the information about the original physical parameters. For example, pairwise distances between objects in the original parameter space were strongly correlated ($\rho > 0.8$) with pairwise distances in the C2-derived feature space. In addition, we found individual C2 units whose activities were highly correlated with one

[4] For brevity, the models with 5, 7, and 10 stored exemplars were withheld from Table 2, since our analysis revealed these data to merely continue the trends seen with 1, 2, and 3 stored exemplars.

Table 2
Goodness-of-fit of the models tested in Experiment 2; see also Table 3 for further discussion of the models' qualitative properties

	RXM⟨1⟩	RXM⟨2⟩	RXM⟨3⟩	SPC⟨1⟩	SPC⟨2⟩	SPC⟨3⟩	GCM	PBI	WPSM
%-variance [orig]	89.36*	90.98*	91.49	89.36*	90.83*	**91.64**	86.84*	87.10*	84.90*
$-\ln L$ [orig]	75.72*	72.06*	71.32*	75.72*	71.65*	**69.92**	83.41*	83.66*	88.79*
AIC [orig]	**173.44**	178.13*	188.64*	**173.44**	177.30*	185.84*	178.81*	177.32	189.57*
%-variance [HMAX]	80.96*	83.98*	85.00*	80.96*	84.54*	**85.99**	75.62*	78.57*	72.57*
$-\ln L$ [HMAX]	91.24*	84.38*	81.89*	91.24*	82.77*	**78.11**	111.92*	97.70*	118.19*
AIC [HMAX]	204.48*	202.76*	209.78*	204.48*	**199.55**	202.23*	235.85*	205.40*	248.37*

%-variance: larger value indicated better fit.
$-\ln L$, minus loglikelihood: smaller value indicates better fit.
AIC, Akaike information criterion: smaller value indicates better fit.
orig: models were fitted using objects represented by the original stimulus parameters, as in Experiment 1.
HMAX: models were fitted using objects represented by features derived from a feed-forward early-vision network.
Bold numbers: model(s) which gave the best fit in each row.
*Models whose fits were significantly worse ($p < 0.05$) than the best-fitting model in each row.

of the original physical parameter values (EH: $\rho = 0.93$, ES: $\rho = 0.91$, NL: $\rho = 0.95$, MH: $\rho = 0.998$).

Among the HMAX-based models, the SPC and RXM again gave better fits than the other models (see Table 2, rows 4–6). As before, the uncorrected measures (minus loglikelihood and %-variance) improved as the number of stored exemplars increased, with the best overall fit given by the SPC⟨3⟩. In contrast to the fits based on the physical parameters, the best AIC values were obtained with two (rather than one) stored exemplars per category for both the SPC and RXM, although as before fits decreased again with more than two stored exemplars. Overall, the HMAX-based model fits were significantly poorer than the corresponding fits based on the physical parameters. Nevertheless, the absolute difference between the best-fitting HMAX-based and physical parameter-based models was only 5.6 %-variance.

4. Discussion

Several authors (Edelman, 1999; Shepard, 1987) have proposed that neural mechanisms of representation are based on similarity. Similarity measures can be transformed to feature space representations with multidimensional scaling, a technique that has often been used as the basis for models of categorization and recognition (e.g., Nosofsky, 1986). Yet, only recently has the neurobiological validity of MDS begun to be investigated directly with monkey electrophysiology (Op de Beeck et al., 2001; Sigala & Logothetis, 2002) and human fMRI (Edelman et al., 1998). Given the practical significance of comparing results obtained in monkey and human studies, it is important to establish the compatibility of the behavioral methods used for the two species. Because it is impossible for monkey observers (as well as for human infants; e.g., Arabie et al., 1975; Sloutsky & Lo, 1999) to give an analog similarity rating, a task based on binary choice such as the "triads" task must be used instead. [5] Unfortunately, since each triads trial conveys only relative information about pairwise similarities, the entire task requires many trials and is quite time demanding. Thus, adult human subjects prefer the "pairs" task, which is based on analog similarity judgments, and is less time demanding since each trial directly conveys absolute information about pairwise similarities. Therefore, we compared the results of the pairs and triads tasks within a set of human subjects to assess their equivalence in characterizing psychophysical representations of similarity. As Fig. 5b shows, the judgments obtained in theses two tasks were highly correlated, suggesting that a shared process could account for subjects' performance in both tasks. These results legitimize comparisons between data from the pairs task in human subjects and data from the triads task in monkey subjects.

One purpose of the MDS analysis is to construct an input representation for the categorization models that can be tested independently of the original stimulus configuration. We found that model fits did not improve when the models were based on pairs-MDS or triads-MDS configurations, relative to the original stimulus configuration. This result agrees with the findings of Sigala et al. (2002) using both monkey and human subjects in experiments similar to those reported here. Thus, although some models (such as the GCM; Nosofsky, 1986, 1991) have originally been used exclusively with MDS configurations, we found that they achieve similar performance when the original configuration is used instead. We interpret these results to mean that subjects can efficiently learn a psychological representation that is highly similar to the native representation of a set of objects. The mechanism for this learning process remains a subject for future investigation. Nevertheless, the empirical correlation between the

[5] Alternatively, a same/different task can be used to generate a confusion matrix for MDS (Sugihara, Edelman, & Tanaka, 1998).

original and MDS configurations is of practical relevance because the MDS procedure is time-intensive both in the collection of similarity task data and in the computational analysis of those data. Our results suggest that this analysis step can be bypassed without affecting the comparison of various classification models.

Experiment 1 revealed a pattern of model fits similar to that reported previously (e.g. Maddox & Ashby, 1993; Nosofsky, 1991; Reed, 1972; Sigala et al., 2002). We found that across several categorization tasks involving different types of objects, an all-exemplar model provided better fits than a linear boundary model, prototype model, or cue-validity model (Table 1). In some cases the fits of the linear boundary and prototype models approached those of the all-exemplar model.

The relative strengths of all-exemplar models and boundary models have been discussed at length (Maddox & Ashby, 1998; McKinley & Nosofsky, 1996; Nosofsky, 1998). Since each model differs from the others in more than one way, it is difficult to conclude which of these differences contribute to a model's success under particular test conditions. To address this point, we introduced a "roaming-exemplar" model (RXM) that can treat independently some of the factors that were mutually dependent in previous models. It shares a flexible memory storage architecture with the striatal pattern classifier (Ashby & Waldron, 1999; Ashby, Waldron, Lee, & Berkman, 2001). It shares a decision mechanism with all-exemplar models and prototype models, since new exemplars are classified by comparing the sums of their similarities to the stored exemplars associated with each of two categories. However, in the roaming-exemplar model as well as the striatal pattern classifier, these stored exemplars are not strictly determined by the training exemplars, but are allowed to "roam" during training within the feature space of the objects to be classified.

In Experiment 2, we analyzed individual subjects' categorizations of 12 different sets of Brunswik faces by fitting them with the roaming-exemplar model and striatal pattern classifier, in addition to the models used in Experiment 1 (Table 2). While the relationships among the all-exemplar, prototype, and linear boundary models have been analyzed previously (Ashby & Alfonso-Reese, 1995; Ashby & Maddox, 1993; Nosofsky, 1990), the improved model fits obtained with the RXM and SPC in Experiment 2 afford new insights into the strengths and weaknesses of previous models (see Table 3 for an overview).

All-exemplar vs. prototype models. There are two significant differences between these models. First, in prototype models, the stored exemplars are by construction defined as the arithmetic mean in feature space of the training exemplars, while in all-exemplar models the stored exemplars occupy other locations. Second, all-exemplar models allow more than one stored exemplar per category, while prototype models allow only one, regardless of the number of training exemplars.

This second difference is linked with the question of category abstraction: storage of a category prototype implies a more abstract representation than simple memorization of all training exemplars. This places a higher burden on the learning process, since the system must select the *correct* abstraction, but makes post-learning categorization more simple, since new exemplars have only to be compared with the category prototypes. All-exemplar models make the opposite trade-off: since no abstraction is involved, learning is straightforward as each training exemplar is simply packed away into memory, but post-learning categorization is complicated since a new exemplar must be compared with every stored exemplar in memory. While this requirement is not neurobiologically unreasonable in typical psychophysical experiments which use few training exemplars per category, it seems less likely to be applicable to natural visual categories, which may contain thousands or more of exemplars. Furthermore, biological systems are likely to spend more time in using categories than in learning them, at least for highly salient categories (e.g., male/female faces, poisonous/non-poisonous fruit). Such arguments lend some a priori credence to the notion of a prototype model, but are entirely hidden from statistical comparisons, since neither the *contents* of the memory nor the complexity of the learning process are free parameters of the models. Indeed, past comparisons between all-exemplar and

Table 3
Qualitative comparison of the key models that were tested in Experiment 2

Model type	Stored exemplars	Main decision boundary		Iso-probability contours	Goodness-of-fit
		Shape	Orientation		Rank (AIC)
Linear boundary	None	Linear	Arbitrary	Linear	2 (177.3)
Prototype	1, fixed	Linear	Constrained	Curved	4 (189.6)
Roaming-exemplar⟨1⟩	1, "roaming"	Linear	Arbitrary	Curved	1 (173.4)
Striatal-pattern⟨1⟩	1, "roaming"	Piecewise-linear	Arbitrary	Piecewise-linear	1 (173.4)
All-exemplar	N, fixed	Curved	Constrained	Curved	3 (178.7)
Roaming-exemplar⟨N⟩	N, "roaming"	Curved	Arbitrary	Curved	5 (279.8)

N: number of training exemplars per category.
AIC, Akaike information criterion: smaller value indicates better fit.

prototype models have generated a preponderance of evidence favoring the all-exemplar models.

When the contents of the memory locations become explicit free parameters, questions concerning the importance of memory capacity can be addressed statistically. For example, by comparing either the RXM$\langle 1\rangle$ or the equivalent SPC$\langle 1\rangle$ with a prototype model, we examine only the first difference mentioned above between all-exemplar models and prototype models (whether memory traces are fixed at the category mean). On the other hand, by comparing the RXM$\langle 1\rangle$ with the RXM$\langle n\rangle$ ($n > 1$) we examine only the second difference (changing the number of stored exemplars). Our results from Experiment 2 (Table 2) demonstrate a large improvement from allowing roaming, rather than fixed, stored exemplars (AIC: RXM$\langle 1\rangle$, SPC$\langle 1\rangle = 173.4$, prototype $= 189.6$), while allowing additional stored exemplars actually leads to a decline in goodness-of-fit when the additional memory is counted among the models' free parameters (AIC: RXM$\langle 10\rangle = 271.9$, RXM$\langle 1\rangle = 173.4$). Thus, although the empirical success of all-exemplar models appears to support a rejection of category abstraction, our results show that in fact we should only reject the strict notion of abstraction involving category prototypes.

Prototype vs. linear boundary models. These two models are similar in that each has a *decision boundary* (i.e., the iso-probability density surface where the categorization probability density equals 0.5) that is a hyperplane in stimulus parameter space (Ashby & Maddox, 1993). The models also have two important differences. First, for prototype models, the decision boundary must be orthogonal to the vector connecting the two category prototypes in stimulus parameter space, while for linear boundary models, the decision boundary can have an arbitrary orientation. Second, consider the iso-probability density surfaces with $p \neq 0.5$: for the linear boundary model, these are hyperplanes parallel to the decision boundary, but for the prototype model, these are paraboloid surfaces with a curvature that increases as p diverges from 0.5. Conceptually, this means that for the linear boundary model, decision thresholds are the same at every point along the category boundary in feature space, while for the prototype model, decision thresholds are narrowest (i.e., the model is most confident) at the center of feature space, near the category prototypes. Intuitively, the behavior of the prototype model seems more natural—new objects are categorized more accurately when they are similar to previously seen objects—but our results from Experiment 1 along with others' results (e.g., Nosofsky, 1991) clearly contradict this intuition.

Again, a more flexible model can help to provide some insight into this issue. In particular, the RXM$\langle 1\rangle$ and SPC$\langle 1\rangle$ are like the prototype model with curved, rather than planar, iso-probability surfaces, but are like the linear boundary model in that the main decision boundary can have an arbitrary orientation. Our results from Experiment 2 demonstrate that with these two qualities combined, the RXM$\langle 1\rangle$ and SPC$\langle 1\rangle$ fit human behavior significantly better than either the prototype or linear boundary models (AIC: RXM$\langle 1\rangle$, SPC$\langle 1\rangle = 173.4$, prototype $= 189.6$, linear boundary $= 177.3$).

All-exemplar vs. linear boundary models. By extension of the previous two comparisons, the differences between the all-exemplar model and the linear boundary model are even more numerous. The all-exemplar model allows for curved decision surfaces, but the orientation of the surface has limited flexibility. In contrast, the linear boundary model allows only flat decision surfaces, but these may have arbitrary orientation. Again, the RXM can combine the separate strengths of these two models.

In the RXM, the parameters which describe the stored exemplars become free parameters of the model, and can be incorporated into comparisons among models using statistical measures such as the Akaike information criterion. This allows us to address the importance of memory by comparing different versions of the RXM with different numbers of stored exemplars. With this framework, we can now provide a better answer as to why models which are otherwise appealing in their conceptual simplicity, such as prototype models, are consistently outperformed by all-exemplar models: all-exemplar models allow better flexibility in matching the shape and orientation of decision surfaces to those used by human observers. Our results show that the goodness-of-fit of all-exemplar models can be improved upon by allowing "roaming" stored exemplars, and thus an unconstrained decision boundary, without committing to high memory demands or to a lack of category-level abstraction.

RXM vs. SPC. Computationally, the RXM and SPC are quite similar to each other, as well as to several earlier models (Anderson, 1991; Kruschke, 1992; see also Ashby & Waldron, 1999), in that they each rely on a set of units representing locations in feature space, and categorize new inputs based on the distance in feature space between the input and the various stored units. The main qualitative difference is at the decision stage, where the RXM produces smoothly curved decision boundaries, while the SPC produces piecewise-linear decision boundaries. This is because in the RXM, the categorization decision is based on contributions from all of the stored units, with weights proportional to the distance of the stored units from the input, while in the SPC, only the nearest stored unit of each category is considered. In this sense, the SPC involves a much stronger non-linearity than the RXM. This sharp non-linearity may not be strictly implemented in neural circuitry; rather, a biological implementation might have to rely on a "softmax" approximation (Riesenhuber &

Poggio, 1999) which would more closely resemble a gradual decay of similarity with distance as in the RXM. This question remains to be resolved by further neurophysiological study.

Despite the computational similarities of the SPC and RXM, each is derived from previous models whose neurobiological implications may appear to put the two models at odds. We have proposed the RXM as a generalization of prototype models and all-exemplar models (i.e., GCM). All-exemplar models, which propose one hidden unit for every training exemplar, have in particular carried the implication that observers may rely on explicit memory of individual visual stimuli, and that the models' hidden units correspond to these memory traces (e.g., Knowlton, 1999). This stands in contrast to neuropsychological evidence from patients with amnesia who, despite an impairment in recognition tasks requiring declarative memory of individual exemplars, are relatively unimpaired in various tasks requiring category learning (Filoteo, Maddox, & Davis, 2001; Knowlton & Squire, 1993; Squire & Knowlton, 1995). For this reason, Ashby and Waldron (1999) proposed that the striatal units in the SPC are primarily response-associated; that is, the units are primarily involved in decision, rather than perception. We do not make any claims regarding whether the hidden units in the RXM are essentially explicit memory traces, particularly since the hidden units are allowed to occupy points in feature space that were never directly related to a training exemplar. However, electrophysiogical evidence does suggest that the mechanisms that are shaped during category learning also affect perception. Sigala and Logothetis (2002) showed that, after category learning, inferotemporal neurons in the macaque were more sensitive to features that were diagnostic of category membership than to non-diagnostic features (although Ashby & Ell, 2001 reviewed studies in which exposure to visual stimuli that were associated with *non-visual* categories such as good/bad tastes did *not* lead to a change in visual cell response properties). Furthermore, behavioral data (MDS) showed that monkeys' *perception* also shifts as a result of category training (Sigala et al., 2002), supporting the idea that the hidden units tuned to specific features in a categorization model may not operate solely at the decision stage, but may also be directly involved in perception. This is not incompatible with the evidence from amnesic patients; it may be that categorization relies on neural representations that are explicit in the sense of being discrete and minimally distributed, but do not constitute "explicit memory" in the sense of being behaviorally accessible for declarative memory. In any case, current psychophysical evidence alone cannot discriminate whether a model's mathematical constructs correspond to neuronal processes occurring in specific cortical areas such as the striatum, inferotemporal cortex, or even prefrontal cortex.

HMAX. We have begun to ground these high-level models of categorization more firmly in neurobiology by combining them with a model (HMAX; Riesenhuber & Poggio, 1999) that encapsulates the processes that functionally precede object categorization in the visual system. Unlike the original categorization models which receive a high-level feature-based description of their input, these hybrid models operate directly on a pixel-based image space representation of the input. Although the hybrid models fit relatively poorly when compared with the original models, their absolute performance is encouraging. The best-fitting HMAX-SPC$\langle 3 \rangle$ model was able to account for nearly 86% of the variance seen in subjects' responses. If anything, our results underestimate the capabilities of a hybrid model, since we used only the first 4 of 576 principal component vectors of the raw HMAX output, sacrificing $\approx$20% of the available variance. This performance was achieved using straightforward bottom–up processing of the input images, with no task-specific training or context-specific top–down modulation of the early-vision stage. Yet, such top–down effects are certainly involved in the performance of human subjects, and the original high-level features are indeed a close approximation of subjects' internal representations as shown by MDS experiments. It thus appears that current high-level models of categorization can be linked to more detailed biological models of vision. A better integration of early-vision and object-categorization models—for example, by allowing attentional weights to propagate from the decision stage back to earlier sensory levels—is likely to uncover a more complete picture of the categorization process.

Generalization and learning. In the most general terms, categorization is a process with four components: (1) external input (visual stimuli), (2) internal input (pre-existing memories and neural state), and (3) a mechanism that combines the inputs to produce (4) an observable output (categorization behavior). A complete theory of categorization should quantitatively describe an internal mechanism that can be appropriately tuned by a learning process involving exposure to a limited set of training exemplars (e.g., Ashby & Ell, 2001; Nosofsky et al., 1992), and should describe how differences in observers' pre-existing internal states lead to different categorization behavior given the same input. In the context of the RXM or SPC, for example, such a theory might help address questions such as how the number of hidden units is adjusted during learning, perhaps in relation to the difficulty in separating categories from one another.

By this standard, the models we have discussed provide only a partial theory, in that they only describe the fully trained mechanism without offering a process for learning the tunable parameters of that mechanism. We have inferred the final values of these parameters by fitting the models to human behavior on a set of test

exemplars. [6] In other words, by collecting and modeling observers' responses to the test exemplars, we have only addressed the question of *what did observers learn*, rather than the more complex question of *how did they learn it*. Nevertheless, our descriptive results provide valuable constraints for more complete future models of the learning process; after all, a model cannot successfully describe the learning process without also successfully describing the outcome of that process.

An open question is to what extent these computational insights, based on psychophysical experiments using simple, four-feature stimuli, carry over to the identification and categorization of complex objects in natural scenes. One challenge is to translate this analysis of the computational principles underlying object categorization into a mature understanding of how neurons along the ventral visual pathway can implement such operations (Op de Beeck et al., 2001; Sigala & Logothetis, 2002).

Acknowledgements

This work was supported by a Predoctoral Fellowship from the Howard Hughes Medical Institute to R.J. Peters. Additional support was provided by the Engineering Research Centers Program of the National Science Foundation under Award Number EEC-9402726, by the NIMH and by the W.M. Keck Foundation Fund for Discovery in Basic Medical Research at Caltech.

References

Anderson, J. (1991). The adaptive nature of human categorization. *Psychological Review, 98*(3), 409–429.

Arabie, P., Kosslyn, S., & Nelson, K. (1975). A multidimensional scaling study of visual memory of 5-year olds and adults. *Journal of Experimental Child Psychology, 19*, 327–345.

Ashby, F. (Ed.). (1992a). *Multidimensional models of perception and cognition*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Ashby, F. (1992b). *Multidimensional models of perception and cognition*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Ashby, F., & Alfonso-Reese, L. (1995). Categorization as probability density-estimation. *Journal of Mathematical Psychology, 39*(2), 216–233.

Ashby, F., & Ell, S. (2001). The neurobiology of human category learning. *Trends in Cognitive Sciences, 5*(5), 204–210.

Ashby, F., & Gott, R. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*, 33–53.

Ashby, F., & Maddox, W. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology, 37*, 372–400.

Ashby, F., & Waldron, E. (1999). On the nature of implicit categorization. *Psychonomic Bulletin & Review, 6*(3), 363–378.

Ashby, F., Waldron, E., Lee, W., & Berkman, A. (2001). Suboptimality in human categorization and identification. *Journal of Experimental Psychology-General, 130*(1), 77–96.

Borg, I., & Groenen, P. (1997). *Modern multidimensional scaling: Theory and applications*. New York: Springer-Verlag.

Brunswik, E., & Reiter, L. (1937). Eindruckscharaktere schematisierter gesichter. *Zeitschrift fuer Psychologie, 142*, 67–134.

Collett, D. (1991). *Modelling binary data*. Boca Raton: Chapman & Hall/CRC.

Edelman, S. (1999). *Representation and recognition in vision*. Cambridge, MA: MIT Press.

Edelman, S., Grill-Spector, K., Kushnir, T., & Malach, R. (1998). Toward direct visualization of the internal shape representation space by fMRI. *Psychobiology, 26*(4), 309–321.

Filoteo, J., Maddox, W., & Davis, J. (2001). Quantitative modeling of category learning in amnesic patients. *Journal of the International Neuropsychological Society, 7*(1), 1–19.

Freedman, D., Riesenhuber, M., Poggio, T., & Miller, E. (2001). Categorical representation of visual stimuli in the primate prefrontal cortex. *Science, 291*(5502), 312–316.

Hernstein, R., & de Villiers, P. (1980). Fish as a natural category for people and pigeons. *The Psychology of Learning and Motivation, 14*, 59–95.

Huber, L., & Lenz, R. (1996). Categorization of prototypical stimulus classes by pigeons. *Quarterly Journal of Experimental Psychology Section B: Comparative and Physiological Psychology, 49*(2), 111–133.

Ingber, L. (1989). Very fast simulated re-annealing. *Mathematical and Computer Modelling, 12*(8), 967–973.

Ishai, A., Ungerleider, L., Martin, A., Schouten, H., & Haxby, J. (1999). Distributed representation of objects in the human ventral visual pathway. *Proceedings of The National Academy of Sciences of The United States of America, 96*(16), 9379–9384.

Jovicich, J., Peters, R., Koch, C., Chang, C., & Ernst, T. (2000). Human perception of faces and face cartoons: An fMRI study. In *Annual meeting of the International Society for Magnetic Resonance in Medicine*, 2000.

Kanwisher, N., McDermott, J., & Chun, M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience, 17*(11), 4302–4311.

Knowlton, B. (1999). What can neuropsychology tell us about category learning? *Trends in Cognitive Sciences, 3*(4), 123–124.

Knowlton, B., & Squire, L. (1993). The learning of categories—parallel brain systems for item memory and category knowledge. *Science, 262*(5140), 1747–1749.

Kruschke, J. (1992). ALCOVE—an exemplar-based connectionist model of category learning. *Psychological Review, 99*(1), 22–44.

Maddox, W. (1999). On the dangers of averaging across observers when comparing decision bound models and generalized context models of categorization. *Perception & Psychophysics, 61*(2), 354–374.

Maddox, W., & Ashby, F. (1993). Comparing decision bound and exemplar models of categorization. *Perception & Psychophysics, 53*(1), 49–70.

Maddox, W., & Ashby, F. (1998). Selective attention and the formation of linear decision boundaries: Comment on McKinley and Nosofsky (1996). *Journal of Experimental Psychology: Human Perception and Performance, 24*(1), 301–321.

[6] An alternate approach would be to fit the models to match observers' performance on the training set, and then judge the models based on observers' performance on the test set. But our observers were trained to be highly accurate in categorizing the training set (with most categorization probabilities near 0 or 1), so their training-set performance places only very weak constraints on the models, since all of the models can be trivially fitted to classify the test set with 100% accuracy. In contrast, we designed the test exemplars for the express purpose of being potentially ambiguous, so that observers' test-set performance would place strong constraints on the models being fitted.

McKinley, S., & Nosofsky, R. (1996). Selective attention and the formation of linear decision boundaries. *Journal of Experimental Psychology: Human Perception and Performance, 22*(2), 294–317.

Nelder, J., & Mead, R. (1965). A simplex method for function minimization. *Computer Journal, 7*, 308–313.

Nosofsky, R. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*(1), 104–114.

Nosofsky, R. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General, 115*(1), 39–57.

Nosofsky, R. (1990). Relations between exemplar-similarity and likelihood models of classification. *Journal of Mathematical Psychology, 34*(4), 393–418.

Nosofsky, R. (1991). Tests of an exemplar model for relating perceptual classification and recognition memory. *Journal of Experimental Psychology: Human Perception and Performance, 17*(1), 3–27.

Nosofsky, R. (1998). Selective attention and the formation of linear decision boundaries: Reply to Maddox and Ashby (1998). *Journal of Experimental Psychology: Human Perception and Performance, 24*(1), 322–339.

Nosofsky, R., Kruschke, J., & McKinley, S. (1992). Combining exemplar-based category representations and connectionist learning rules. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*(2), 211–233.

Op de Beeck, H., Wagemans, J., & Vogels, R. (2001). Inferotemporal neurons represent low-dimensional configurations of parameterized shapes. *Nature Neuroscience, 4*(12), 1244–1252.

Poggio, T., & Girosi, F. (1990). Regularization algorithms for learning that are equivalent to multilayer networks. *Science, 247*(4945), 978–982.

Reed, S. (1972). Pattern recognition and categorization. *Cognitive Psychology, 3*, 382–407.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience, 2*(11), 1019–1025.

Rosseel, Y. (1996). Connectionist models of categorization: A statistical interpretation. *Psychologica Belgica, 36*(1–2), 93–112.

Shepard, R. (1987). Toward a universal law of generalization for psychological science. *Science, 237*(4820), 1317–1323.

Sigala, N., Gabbiani, F., & Logothetis, N. (2002). Visual categorization and object representation in monkeys and humans. *Journal of Cognitive Neuroscience, 14*(2), 187–198.

Sigala, N., & Logothetis, N. (2002). Visual categorization shapes feature selectivity in the primate temporal cortex. *Nature, 415*(6869), 318–320.

Sloutsky, V., & Lo, Y. (1999). How much does a shared name make things similar? Part 1: Linguistic labels and the development of similarity judgment. *Developmental Psychology, 35*(6), 1478–1492.

Smith, J., & Minda, J. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*(6), 1411–1436.

Squire, L., & Knowlton, B. (1995). Learning about categories in the absence of memory. *Proceedings of The National Academy of Sciences of The United States of America, 92*(26), 12470–12474.

Sugihara, T., Edelman, S., & Tanaka, K. (1998). Representation of objective similarity among three-dimensional shapes in the monkey. *Biological Cybernetics, 78*(1), 1–7.

Vogels, R. (1999). Categorization of complex visual images by rhesus monkeys. Part 1: behavioural study. *European Journal of Neuroscience, 11*(4), 1223–1238.

Zucchini, W. (2000). An introduction to model selection. *Journal of Mathematical Psychology, 44*(1), 41–61.